Metadaten und die Data Documentation Initiative (DDI)

Zenk-Möltgen, Wolfgang

Veröffentlichungsversion / Published Version
Sammelwerksbeitrag / collection article

Metadaten und die Data Documentation Initiative (DDI)

Wolfgang Zenk-Möltgen

Über DDI

Die Data Documentation Initiative (DDI)[1] ist eine Initiative, einen internationalen Standard zur Beschreibung sozialwissenschaftlicher Daten zu definieren und zu verbreiten. Dieser Standard wird in einem XML-Format (Extensible Markup Language) definiert, das sowohl für Menschen als auch für Maschinen lesbar ist. DDI hat den Anspruch, den gesamten Forschungsdaten-Lebenszyklus zu unterstützen. DDI Metadaten beziehen sich auf die Studienkonzeption, die Datenerhebung, die Datenbearbeitung und -auswertung sowie auf die Sekundärnutzung und Archivierung.

Die Konzeption und Definition der Ziele der Initiative kamen aus der Welt sozialwissenschaftlicher Datenarchive. 1995 wurde DDI als Projekt finanziert, gestartet und organisiert vom ICPSR (Inter-university Consortium for Political and Social Research, USA).[2] 2003 wurde die DDI Alliance gegründet, welche auf der Mitgliedschaft von Institutionen basiert und formalisierte Prozesse zur Weiterentwicklung der Initiative einführte. Die Gründungsmitglieder kamen aus dem Bereich sozialwissenschaftlicher Datenarchive, den Produzenten von Statistikdaten und weiteren, wie zum Beispiel von Forschungsdatenzentren, Datenerhebungsinstitutionen und einigen kommerziellen Organisationen. Heute sind 36 Institutionen aus 14 Ländern in Nordamerika, Europa und Australien Mitglieder in der DDI Alliance, zusätzlich Eurostat und die World Bank Development Data Group als internationale Organisationen (siehe Tabelle 1). Der Bereich der Anwender von DDI ist noch weiter als der der Mitglieder von DDI. Allein die World Bank setzt DDI in über 100 Statistik-Ämtern in 67 Ländern ein. Auf der Website der DDI Alliance wird mithilfe der *„DDI is being used around the world"* die weltweite Verbreitung von DDI gut dargestellt.[3]

University of Alberta, Canada

Australian Bureau of Statistics (ABS)

Australian Data Archive (ADA)

1 http://www.ddialliance.org
2 http://www.icpsr.umich.edu
3 http://www.ddialliance.org/community

University of California, Berkeley - Computer-Assisted Survey Methods Program and UCDATA
Centre for Longitudinal Studies, Institute of Education, University of London (Associate Member)
Centro De Investigaciones Sociologicas (CIS), Spain
Cornell University (CISER)
Danish Data Archive
Data Archiving and Networked Services (DANS), The Netherlands
Eurostat
Finnish Social Science Data Archive
German Institute for International Educational Research (DIPF)
German Socio-Economic Panel Study (SOEP)
GESIS - Leibniz Institute for the Social Sciences
University of Guelph
Institute for Quantitative Social Science (IQSS) at Harvard University
Institute for the Study of Labor (IZA)
Institute for Social and Economic Research (ISER)
Inter-university Consortium for Political and Social Research (ICPSR)
Massachusetts Institute of Technology (MIT)
University of Minnesota, Minnesota Population Center
Norwegian Social Science Data Service (NSD)
Open Data Foundation
Princeton University
Research Data Centre of the German Federal Employment Agency, Institute for Employment Research (IAB)
Roper Center
Stanford University
Statistics New Zealand
Survey Research Operations, University of Michigan
Swedish National Data Service (SND)
Swiss Foundation for Research in Social Sciences (FORS)
United Kingdom Data Archive
University of Toronto Scholars Portal
University of Washington, Center for Studies in Demography & Ecology (CSDE)
U.S. Bureau of Labor Statistics (Associate Member)
World Bank, Development Data Group (DECDG)

Tabelle 1: Mitglieder der DDI Alliance

Die Entwicklung und weitere Verbesserung des DDI Standards führte zu der Veröffentlichung von verschiedenen Versionen. Im Jahr 2000 wurde die DDI Version 1.0 veröffentlicht, in welcher einfache Umfragen dokumentiert werden konnten und zum Beispiel nur Mikrodaten, aber keine Aggregatdaten. Im Jahr 2003 erschienen die Versionen 2.0 und 2.1 als Erweiterung von DDI, in denen nun auch Aggregatdaten und weitere Datentypen dokumentiert werden können sowie eine Unterstützung für geographische Elemente möglich ist. Diese Version von DDI wird auch gegenwärtig noch verwendet und gepflegt und ist unter dem Namen *„DDI-Codebook"* bekannt, da sich die Dokumentation sehr stark an der Struktur eines Codebuchs für einen Datensatz orientiert.

Im Jahr 2008 erschien mit DDI 3.0 die erste *„DDI-Lifecycle"* Version, in der eine Erfassung des Daten-Lebenszyklus im Gegensatz zum Codebuch-zentrierten Modell im Mittelpunkt steht. Hier wurde der Blick auf die Erzeugung der Metadaten und ihre Wiederverwendung in den verschiedenen Stadien des Forschungsdaten-Lebenszyklus gerichtet. Zusätzlich wurde das Konzept eingeführt, dass die Metadaten-Elemente *"machine-actionable"* sein sollten, also so strukturiert, dass ein programmierter Zugriff auf sie möglich ist. Eine Erweiterung der Fragebogendokumentation wurde zur Unterstützung der Verwendung von CAI-Instrumenten (Computer Aided Interview) eingeführt. Weitere Neuerungen betrafen die Unterstützung für Datenreihen (Längsschnitt-Umfragen, Panel Studien, etc.), die Möglichkeit zum Vergleich von Metadaten (*"by design"* und *"ex-post"* möglich) und eine verbesserte Unterstützung zur Beschreibung komplexer Datensätze.

2009 wurden mit der Version DDI 3.1 etliche Fehlerkorrekturen durchgeführt, darunter auch eine neue URN-Struktur, um dauerhafte Identifikatoren aller "identifiable"-Elemente von DDI zu erhalten (siehe unten). Die Version 3.1 ist gegenwärtig die aktuelle Version von DDI-Lifecycle.

Als Update zur DDI-Codebook Variante wurde in 2012 die Version DDI 2.5 veröffentlicht. Sie soll eine Erleichterung der Migration von DDI-Codebook nach DDI-Lifecycle ermöglichen, indem zum Beispiel alle notwendigen Elemente (Pflichtelemente) von DDI-Lifecycle mit einem Gegenstück in DDI 2.5 eingeführt wurden. Dies erleichtert vor allem auch die parallele Verwendung von DDI-Codebook und DDI-Lifecycle, da die Pflichtelemente von DDI-Lifecycle bei einer Konvertierung in das DDI-Codebook Format nicht verloren gehen.

Für 2012 ist eine Veröffentlichung von DDI 3.2 angekündigt. Darin soll zum Beispiel ein Element *"DataItem"* neu eingeführt werden, das eine Wiederverwendung erlaubt, und es sollen einige Konsistenz-Fragen gelöst werden, etwa bei *"RecordRelationship"* oder bei der Verwendung von Missing Values. Weiterhin wird die URN-Struktur überarbeitet, damit ein verteilter Resolving-Mechanismus für DDI-URNs möglich wird. Zusätzlich wird die Verwendung kontrollierter Vokabulare verbessert werden.

Grundlegende DDI Metadaten

Die DDI-Codebook Version enthält Metadaten zu den vier Bereichen: Dokumentbeschreibung, Studienbeschreibung, Variablenbeschreibung und Dateibeschreibung. Die wichtigsten Elemente innerhalb dieser Bereiche sind für die Dokumentbeschreibung der Titel, die Autoren und die Beschreibung der Publikation des DDI Dokuments selbst. Für die Studienbeschreibung sind vor allem wichtig die Inhalte der Studie, Titel, Autoren und Institutionen, zeitliche und geographische Angaben zur Studie, verwendete Methoden, Grundgesamtheit und Stichprobenbeschreibung sowie Literaturhinweise. Der Bereich Variablenbeschreibung enthält als Hauptelemente die Namen, Typen und Labels zu den Variablen, die Fragen und Antwortmöglichkeiten der Interviews, verwendete Codes und ihre Häufigkeiten im Datensatz, Interviewer-Anweisungen und Filterinformationen aus dem Fragebogen und Hinweise zur Codierung oder Berechnung. Die Dateibeschreibung schließlich enthält Angaben zur Anzahl der Variablen und Fälle und Namen, Formate und Versionen der Datendateien.

Die DDI-Lifecycle Version folgt hier einem anderen Konzept, nämlich einer unabhängigen Dokumentation einzelner Stadien des Lebenszyklus und damit der Möglichkeit ihrer Wiederbenutzung. Im Folgenden werden daher die Prinzipien der Strukturierung von DDI-Lifecycle näher erläutert: der Lebenszyklus, DDI Module, Elemente, die als Maintainables, Versionables und Identifiables klassifiziert werden können, die Einführung von Schemes (pflegbare Listen). Desweiteren wird auf die Beziehungen zu anderen Standards eingegangen und es wird die Verwendung kontrollierter Vokabulare in DDI 3 genauer erläutert.

Forschungsdaten-Lebenszyklus

Die DDI Alliance hat einen Lebenszyklus für Forschungsdaten definiert, in dem die verschiedenen Phasen als Struktur für die verwendeten Module in DDI 3 dienen können (siehe Abbildung 1).

Als Phasen wurden dabei acht verschiedene identifiziert, die jedoch nicht in linearer Reihenfolge durchlaufen werden müssen. Mit der Phase Concept beginnt eine Studienkonzeption, indem Forschungsfrage und Methodik der Untersuchung festgelegt werden. Anschließend wird in der Phase Collection die Datenerhebung durchgeführt und in Phase Processing die Datenaufbereitung geleistet. Von hier ab kann entweder zur Phase Archiving weitergegangen werden, in der eine Sicherung der Forschungsdaten geleistet wird, oder aber auch direkt zur nachfolgenden Phase Distribution, welche den Datenvertrieb leistet. Von hier folgt eine Phase Discovery, die die Daten auffindbar macht, und die Phase Analysis, in der die Forschungsdaten ausgewertet werden. Hier schließt sich dann noch eine Phase Repurposing an, die eine anderweitige Verwendung

der Daten für Sekundärnutzung umfasst und dann wieder zur Phase Processing zurückführt. Dieses allgemeine Modell der Phasen des Forschungsdaten-Lebenszyklus diente als Grundlage für die Entwicklung der Module. Die Reihenfolge der Verwendung der einzelnen Phasen ist dabei aber komplett offen und nicht in der DDI Spezifikation vorgegeben. Daher sind alle anderen Pfade durch die Phasen denkbar und können in DDI auch so dokumentiert werden.

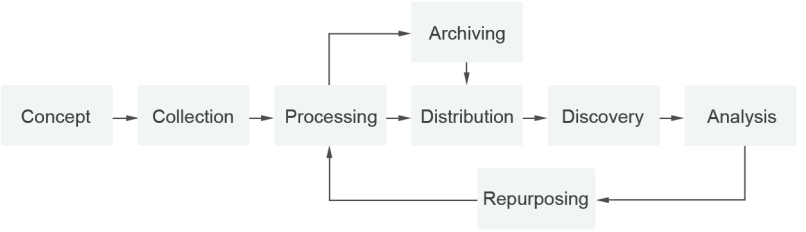


Abbildung 1: Der Forschungsdaten-Lebenszyklus der DDI Alliance

Module

Module in DDI-Lifecycle sind Gruppen von zusammengehörigen Dokumentationselementen. Manche Module beziehen sich auf das Lebenszyklusmodell, andere sind aus technischen Gründen gruppiert. Zusammengehörige Elemente sollten in einem Modul zu finden sein, was aber nicht immer möglich war. Dies hat jedoch keine Auswirkungen auf die Verwendung der Elemente, da die Gruppierung in Modulen allein der internen Strukturierung der DDI Spezifikation dient. Folgende Module sind in DDI 3.1 definiert:

DDI 3.1 Module
Archive module
Comparative module
Conceptual components module
Data collection module
Dataset module
Dublin Core Elements module
DDI profile module
Grouping module
Instance module
Logical product module
Physical data product module *(plus inline n-cube, normal n-cube, tabular n-cube module and proprietary module)*

Physical instance module
Reusable module
Study unit module

Tabelle 2: DDI 3.1 Module

Benutzung der DDI 3 Module

Für einige Module soll hier gezeigt werden, welche Metadaten-Elemente sie enthalten und wie sie verwendet werden (siehe Abbildung 2). Zunächst gibt es das Modul StudyUnit, in dem grundlegende Metadaten über eine einfache Studie enthalten sind. Hier werden zum Beispiel Informationen abgelegt, die zur Identifizierung dienen, etwa Studiennummer, Persistent Identifier oder Zitationsinformationen. Desweiteren sind hier die räumliche und zeitliche Einordnung der Studie und die abgedeckten Themen dokumentiert. Grundlegende Konzepte, die abzubildende Grundgesamtheit, eine inhaltliche Zusammenfassung und Informationen über den Zweck der Studie sowie über Forschungsanträge und ihre Finanzierung sind hier ebenfalls Gegenstand der Dokumentation.

Study Unit	**Data Collection**
Identification Coverage • *Topical* • *Temporal* • *Spatial* Conceptual Components • *Universe* • *Concept* • *Representation (optional replication)* Purpose, Abstract, Proposal, Funding	Methodology Question Scheme • *Question* • *Reponse domain* Instrument • *using Control Construct Scheme* Coding Instructions • *question to raw data* • *raw data to public file* Interviewer Instructions

Logical Product	**Archive**
Category Schemes Coding Schemes Variables NCubes Variable and NCube Groups Data Relationships	Organization or individual which has control over the metdata Livecycle events Archive specific information

Physical Data Structure	Physical Instance
Links to Data Relationships	One-to-one relationship with data file
Links to Variable or NCube Coordinate	Coverage constraints
Description of phsycial storage structure	Variable and category statistics
• *in-line, fixed, delimited or proprietary*	

Abbildung 2: Verwendung der DDI Module

Das nächste Modul Data Collection enthält Informationen zur angewandten Datenerhebungsmethode, dem Instrument – etwa dem Fragebogen oder auch anderen Messinstrumenten – mit den zugehörigen Fragen und Antwortdomänen sowie ihrer Abfolge im Fragebogen. Zusätzlich sind Interviewerанweisungen aus dem Fragebogen und Codierungsanweisungen für die Rohdaten und auch für die letztlich publizierten Datensätze Teil der Metadaten in diesem Modul.

Im Modul Logical Product werden die Metadaten zur Struktur der erhobenen Daten abgelegt: Hier sind die Listen der Antwortkategorien und der verwendeten numerischen Codes und die daraus entstehenden Variablen des Datensatzes dokumentiert. Dazu gehören auch sog. NCubes, das sind aggregierte Daten von Variablen mit mehreren (N) Dimensionen oder generell n-dimensionale Datenstrukturen. Variablen und NCubes können in Gruppen zusammengefasst und dokumentiert werden. Beziehungen zwischen ihnen können ebenfalls beschrieben werden.

Im Modul Physical Data Structure werden die physikalischen Eigenschaften der verwendeten Datenstrukturen dokumentiert, etwa ein festes, ein variables oder ein Trennzeichen-Format. Die Verbindung zu den Variablen aus dem Logical Product erfolgt über die Data Relationships. Die eigentliche Datendatei wird dann im Modul Physical Instance beschrieben. Dort besteht eine Eins-zu-eins-Relation mit einer Datei, die die Umfragedaten enthält, etwa einer SPSS- oder STATA-Datei. In diesem Modul können auch Tabellen mit statistischen Ergebnissen zu den Variablen abgelegt werden.

Das Modul Archive ist in einem sehr weiten Sinne zu verstehen, zum Beispiel sind in diesem Modul alle Informationen über beteiligte Personen und Institutionen zu finden. Dazu gehören auch die Lifecycle Events, mithilfe derer alle Prozesse aus den verschiedenen Lebenszyklus-Stadien erfasst werden können. Daneben gibt es hier auch Informationen zur Archivierung der Studie und zu den zugehörigen Katalog-Metadaten.

Neben den hier gezeigten grundlegenden Modulen zur Beschreibung von Forschungsdaten gibt es noch weitere Möglichkeiten in anderen Modulen, etwa zur Dokumentation von Konzepten im Conceptual Components Modul, zum Vergleich verschiedener Elemente im Modul Comparative oder zur Vererbung von Dokumentation durch die Benutzung des Group Moduls. Wichtig ist auch das

Element ResourcePackage aus dem Modul Group: Es dient zur Dokumentation wiederverwendbarer Elemente unabhängig von ihrem Einsatz in einer Studie, zum Beispiel für Fragen, Antwortskalen oder Variablen.

Die Elemente aus allen diesen Modulen können vielfältig miteinander vernetzt werden, indem Referenzen auf Elemente benutzt werden, die an anderer Stelle dokumentiert sind. So kann eine maximale Wiederverwendung der Dokumentationsteile in den verschiedenen Stadien des Forschungsdaten-Lebenszyklus erreicht werden.

Maintainables, Versionables und Identifiables

Die Elemente in DDI 3 können folgendermaßen klassifiziert werden: Zunächst gibt es einfache Elemente, die Metadaten für ein Objekt enthalten oder eine Referenz auf ein anderes Element (siehe Abbildung 3). Auf der nächsten Stufe gibt es die sog. Identifiables, welche zusätzlich über eine ID verfügen. Mithilfe der ID können diese Elemente eindeutig identifiziert werden, so dass auf diese Elemente eine Referenz gesetzt werden kann. Dabei gibt es zwei technische Möglichkeiten, diese ID festzulegen, entweder über ein ID-Attribut oder über eine URN (Uniform Resource Name), die den speziellen Vorgaben der DDI Spezifikation folgt (siehe unten).

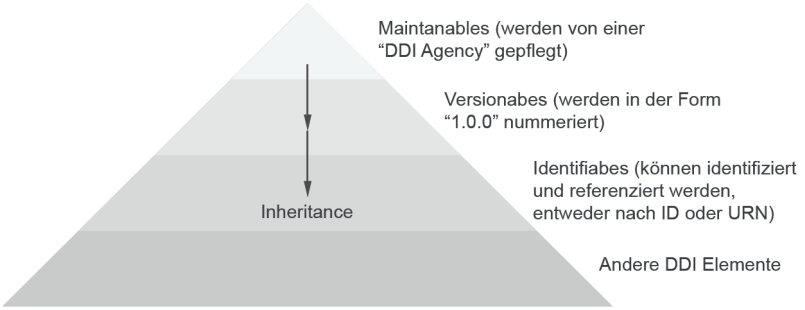


Abbildung 3: Hierarchie der Elemente

Eine weitere Gruppe von Elementen verfügt zusätzlich zur ID über eine Versionsnummer und gehört damit zu den Versionables. Diese Elemente können in verschiedenen Versionen vorliegen, die mithilfe einer dreistelligen Versionsnummer gekennzeichnet wird. Referenzen auf diese Elemente müssen auch die Versionsangabe enthalten.

Die Gruppe der Maintainables schließlich verfügt zusätzlich zur ID und der Version über das Attribut Agency zur Kennzeichnung eines Anbieters, der die Informationen des Elements pflegt. Institutionen können einen Agency-Namen bei der DDI Alliance beantragen und können mit der Verwendung dieses Namens für DDI Instanzen ihre Verantwortlichkeit für die dort enthaltenen Metadaten erklären.

Das Konzept der Schemes

Im DDI Standard stehen Schemes für Listen von Elementen eines gleichen Typs. Um die Verwaltung dieser Elemente zu vereinfachen, können sie zu Schemes zusammengefasst werden. Schemes sind in der Regel Maintainables und können daher von einer DDI Agency gepflegt und von ihr selbst oder anderen per Referenz wiederverwendet werden. Beispiele für Schemes sind das Organization Scheme im Modul Archive, das Question Scheme, Control Construct Scheme und Interviewer Instruction Scheme im Data Collection Modul, das Concept Scheme, Universe Scheme, Geographic Structure Scheme und Geographic Location Scheme im Modul Conceptual Components, das Category Scheme, Code Scheme, Variable Scheme und NCube Scheme im Modul Logical Product, das Physical Structure Scheme und Record Layout Scheme im Physical Data Product Modul.

Beziehungen von DDI zu anderen Standards

In die Entwicklung des DDI Standards sind eine ganze Reihe von Erfahrungen aus anderen Standards mit eingeflossen. Darüber hinaus können einige andere Standards auch direkt in DDI eingebunden werden. Dazu gehört zum Beispiel der Dublin Core Standard[4] zur Dokumentation grundlegender bibliographischer Zitationsinformationen und zur Dokumentation von Sammlungen und vorliegenden Formaten. Die Definition von Dublin Core wurde so eingebunden, dass diese Elemente an bestimmten Stellen direkt innerhalb von DDI verwendet werden können: Dazu dient das DCElements-Tag, das in jedem Citation-Element verwendet werden kann.

Von grundlegendem Einfluss für die Entwicklung von DDI war das OAIS Referenzmodell für Archive. Viele Elemente, die in OAIS genannt sind, finden sich in der DDI Spezifikation wieder. Daneben wurden aber auch Elemente aus METS mit einbezogen, die eine Beschreibung zum Management digitaler Objekte auf einer oberen Ebene darstellen, und aus PREMIS, einem Standard mit spezifischen Strukturen für die digitale Langzeitsicherung.

Die grundlegenden Konzepte für die Metadaten zu geographischen Informationen wurden entlang des ISO-Standards 19115 Geography (FGDC) entwickelt, der Elemente wie Shape, Boundary oder Map Image Dateien und ihre Attribute enthält. Die Struktur der Beschreibung von Konzepten lehnt sich am ISO/IEC-Standard 11179 an. Dieser sieht eine Repräsentation von Metadaten in einer Registratur (Metadata Registry) vor, die auch eine Hierarchie von Konzepten und eine detaillierte Beschreibung der Konzepte enthalten kann.

Für die Modellierung des DDI Standards in XML-Schema wurde auf Erfahrungen mit dem Standard SDMX zurückgegriffen. Dieser wurde für den Austausch und die Dokumentation von statistischen Aggregatdaten, etwa Zeitreihen oder

4 http://dublincore.org/documents/dcmi-terms/

Indikatoren entwickelt und wird bereits sehr verbreitet durch statistische Ämter und andere Statistikproduzenten eingesetzt. Die Verwendung von DDI und SDMX kann durch die ähnliche Ausrichtung und die inhaltliche Nähe gut komplementär erfolgen, was auch durch eine gemeinsame Arbeitsgruppe der Initiativen noch verbessert werden soll.

Verwendung von kontrollierten Vokabularen

Kontrollierte Vokabulare erlauben es, die Einträge von Metadaten-Elementen auf eine Liste von erlaubten Werten einzuschränken. Dadurch wird eine höhere Standardisierung erreicht, als es mit den Einträgen von Texten zur Beschreibung möglich ist. So kann zum Beispiel die Angabe einer Klassifikation mithilfe eines kontrollierten Vokabulars erfolgen. In DDI-Lifecycle ist daher die Verwendung von kontrollierten Vokabularen möglich und wird empfohlen.

Wegen der Fülle der Möglichkeiten der Anwendung von kontrollierten Vokabularen und der Breite der Themen, auf die sie sich erstrecken können, hat man sich entschieden, die kontrollierten Vokabulare nicht als Teil des DDI Standards zu formulieren, sondern als eine Empfehlung. Eine Arbeitsgruppe der DDI Alliance (CVG) hat bereits Empfehlungen zu einer Reihe von kontrollierten Vokabularen veröffentlicht. Darunter sind zum Beispiel Empfehlungen zu den Elementen LifeCycleEvent, Commonality, TimeMethod, ResponseUnit, SoftwarePackage, CharacterSet und AnalysisUnit. Weitere Elemente, an denen gearbeitet wird, sind IntendedFrequency, ModeOfDataCollection, AggregationMethods, DataType, CategoryStatistic, DateCalendar, ContributorRole, PublisherRole und KindOfData.

Am Beispiel von TimeMethod kann man sehen, dass folgende Werte dort enthalten sein können: Longitudinal (Cohort or Trend), Panel (Continuous or Interval), TimeSeries (Continuous or Discrete), CrossSectional, CrossSectionalAdHocFollowUp und Other. Hiermit werden also die Typen des Studiendesigns über Zeit festgelegt, so dass leichter Suchen durchgeführt werden können oder Gruppen von Studien gebildet werden können.

Diese kontrollierten Vokabulare werden im Format Genericode[5] veröffentlicht, eine Spezifikation von OASIS[6] zur Dokumentation und Versionierung von Codelisten. Manche Vokabulare sind auch als Hierarchien angelegt, so dass sie weitere oder engere Begriffe vorsehen. Die kontrollierten Vokabulare der DDI Alliance werden von der CVG gepflegt, so dass neuere Versionen auch auf der Website der DDI Alliance[7] zu finden sein werden.

5 http://genericode.org

6 https://www.oasis-open.org

7 http://www.ddialliance.org/controlled-vocabularies

Identifizieren von Elementen in DDI 3

Alle Identifiables in DDI 3 verfügen über ein ID-Attribut, das dazu dient, Referenzen auf diese Elemente erzeugen zu können. Mithilfe dieser Referenzen können Elemente wiederbenutzt werden, wobei sie nur einmal in DDI dokumentiert werden müssen.

Es gibt zwei technische Möglichkeiten, die ID festzulegen, entweder über das ID-Attribut oder über ein URN-Attribut (Uniform Resource Name), das den speziellen Vorgaben der DDI Spezifikation folgt. Beide Varianten sind logisch gleich und enthalten die gleichen Angaben. Lediglich die Syntax ist unterschiedlich, aber beide Varianten können ineinander überführt werden. So sprechen allein technische Gründe für die Verwendung der einen oder anderen Variante in einer konkreten Implementierung.

Bei Verwendung des ID-Attributs muss ein eindeutiger Identifier in der Form id="A1234" angegeben werden. Die Angabe zur Agency wird aus dem übergeordneten Maintainable, die Angabe zur Version wird aus dem übergeordneten Versionable geerbt. Mit diesen drei Angaben kann dann eine eindeutige Referenz an anderer Stelle auf dieses DDI Element erzeugt werden.

Bei Verwendung des URN-Attributs werden die Angaben zum Identifier, der Agency und der Version in einen String kombiniert, die den Vorgaben der URN-Spezifikation folgt. Dies wird zum Beispiel für eine Variable V100 im VariableScheme ZA1234_VarSch der GESIS für Version 1.0.0 in der Form urn="urn:ddi:de.gesis:VariableScheme.ZA1234_VarSch.1.0.0:Variable.V100.1.0.0" notiert. Diese Variante der Notation wird empfohlen, beide Varianten sind erlaubt.

Für eine effektive Wiederbenutzung von DDI Elementen wird ein Resolver Service benötigt, der diese Identifier so auflöst, dass die Lokation des DDI Elements gefunden werden kann und seine Eigenschaften dann ermittelt werden können. Die DDI Alliance arbeitet zurzeit an der Implementierung eines solchen Services, die auf der Verwendung des DNS-Systems (Domain Name System) beruht.

DDI im GESIS Datenarchiv

Das GESIS Datenarchiv nutzt heute DDI-Codebook für den Workflow bei der Daten- und Metadatenbearbeitung und für die Langzeitarchivierung der zu archivierenden Studien. Da das Datenarchiv in vielen Projekten auch bereits bei der Datenerhebung, bei der Datenaufbereitung und -pflege, und auch generell bei der Datendistribution und -analyse tätig ist, beschränkt sich die Verwendung von DDI nicht auf die Lebenszyklus-Phase der Archivierung im engeren Sinne, sondern die DDI Metadaten werden auch in der Unterstützung von Dokumentation, langfristiger Sicherung, Recherche und Datenservice für Sekundärnutzer und für

die DOI Registrierung über das Angebot da|ra[8] verwendet. DDI-Lifecycle wird gegenwärtig nur für spezielle Anwendungen, wie etwa die Unterstützung von Enhanced Publications (Verbindung von Publikationen zu den dabei benutzten Daten) oder im Projekt STARDAT verwendet, das eine Integration der Archiv-Tools auf der Basis von DDI-Lifecycle beinhaltet. Eine Migration aller Anwendungen und Metadatenbestände nach DDI-Lifecycle wird zurzeit geplant, dabei sind allerdings noch eine Reihe von Hindernissen zu überwinden, etwa eine Einführung einer effektiven Versionskontrolle für einzelne Metadaten-Elemente.

Workflow

Die Abläufe im GESIS Datenarchiv für die Archivierung, Datendokumentation, -bearbeitung, Langzeitsicherung und Distribution werden zum großen Teil durch DDI-Codebook Metadaten unterstützt (siehe Abbildung 4). Die Studienbeschreibungen im Datenbestandskatalog (DBK)[9] werden im webbasierten Programm DBKEdit in einer DDI kompatiblen relationalen Datenbank erstellt und gepflegt. Sie werden nicht nur in DBKSearch publiziert, sondern dienen auch zur Anbindung an die Datenregistrierung da|ra zur Vergabe von persistenten Identifikatoren (DOI Namen) und an das Nesstar-basierte ZACAT-Angebot[10] zur Analyse, Recherche und zum Download von archivierten Studien.

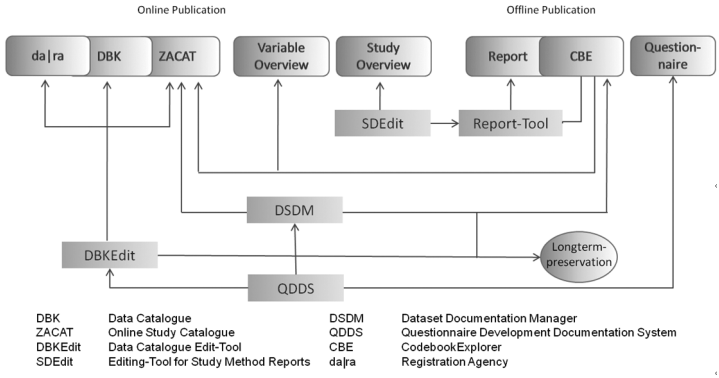


Abbildung 4: Workflow im GESIS Datenarchiv

Die Dokumentation der Studien auf der Variablenebene erfolgt mithilfe der Software Dataset Documentation Manager (DSDM) oder kann auch in der Fragebogensoftware Questionnaire Development and Documentation System (QDDS) erfolgen. DSDM exportiert die Dokumentation in DDI-Codebook Format zur Verwendung in ZACAT oder im CodebookExplorer (CBE), einer stand-alone Software

8 http://www.gesis.org/dara/
9 http://www.gesis.org/dbk/
10 http://zacat.gesis.org

zur Recherche und Analyse von Datenkollektionen. Hierdurch wiederum werden verschiedene Ausgabeformate unterstützt, die alle auf derselben Dokumentationsbasis beruhen: Ein DDI-Codebook Export unterstützt das Nesstar-System, eine CBE-Datenbank kann für einen Variable-Overview benutzt werden, welcher komplexe Datenkollektionen im Web darstellt und durchsuchbar macht, und Variable Reports können als Print- oder Online-Ausgabe für eine vollständige Datensatzdokumentation erzeugt werden.

Die Ausgabe eines speziellen Langzeitsicherungsformats wird sowohl von DBKEdit als auch von DSDM unterstützt. Dieses Format ist vollständig DDI-Codebook kompatibel und wegen seiner Lesbarkeit von Mensch und Maschinen gut als Format für eine langfristige Sicherung geeignet.

Das Beispiel Datenbestandskatalog

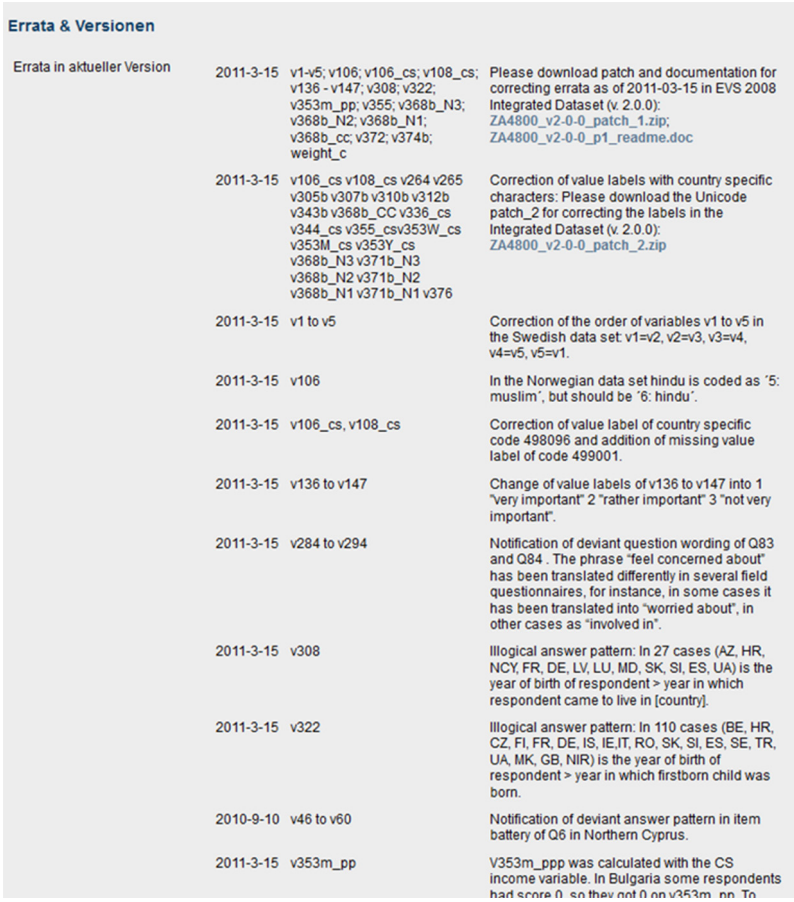
Das GESIS Datenarchiv für Sozialwissenschaften, im Jahr 1960 als Zentralarchiv für empirische Sozialforschung der Universität Köln gegründet, hat bereits seit langer Zeit unter anderem durch die Entwicklung eines standardisierten Studienbeschreibungsschemas an einer Vereinheitlichung von Metadaten gearbeitet. Zusammen mit den anderen Archiven des Verbundes CESSDA (Council of European Social Science Data Archives)[11] wurden diese Bemühungen ständig intensiviert und flossen in die Mitarbeit von vielen CESSDA-Archiven bei der DDI Alliance ein. Die standardisierten Studienbeschreibungen wurden und werden als Datenbestandskatalog (DBK) regelmäßig in gedruckter Form oder online veröffentlicht. Die Anwendung DBKEdit dient seit 2006 als relationales Datenbanksystem zur Verwaltung und Pflege der Studienbeschreibungen, die den Nutzern mithilfe der Anwendung DBKSearch zur Verfügung gestellt werden. DBKEdit leistet dabei auch die Metadaten-Produktion für die Publikation in verschiedenen Retrieval- und Distributionsplattformen (siehe oben). Seit kurzem sind diese Anwendungen als DBKfree[12] unter einer Open Source Lizenz auch für weitere Anwender verfügbar.

Im April 2010 wurde eine Versionshistorie der Datensätze als einheitliches System für alle archivierten Daten mithilfe des DBK eingeführt. Diese enthält eine eindeutige Versionsnummer, eine detaillierte Dokumentation von Errata und der Korrektur-History der Datensätze. Die Errata zur aktuellen Version werden mit Datum, einer Fehlerbeschreibung und einer Beschreibung, welche Variablen betroffen sind, versehen (siehe Abbildung 5). Die Versionsnummern werden DDI-Lifecycle konform mit Version 1.0.0 begonnen und erhöhen sich bei Major-, Minor- oder Revision-Nummer je nach Änderung im Datensatz. Dies führt zu höherer Transparenz im Laufe der Datenbearbeitung. Zusätzlich wird

11 http://www.cessda.org

12 http://info1.gesis.org/dbkfree/

auf diese Art ein einfaches Zitieren der Daten ermöglicht, da eine Version immer mit einem DOI Namen als persistenter Identifikator über das da|ra-System versehen wird. Die Zitation von Datensätzen wird den Nutzern im DBK vorgeschlagen und enthält die genaue Version zur Erleichterung von Replikationsanalysen. Die Studienbeschreibungen zur aktuellen Version des Datensatzes können in DDI-Codebook oder DDI-Lifecycle exportiert werden. Ein Export der Dokumentation der kompletten Versionshistorie im DDI-Format steht allerdings noch aus.



Errata & Versionen

Errata in aktueller Version	2011-3-15	v1-v5; v106; v106_cs; v108_cs; v136 - v147; v308; v322; v353m_pp; v355; v368b_N3; v368b_N2; v368b_N1; v368b_cc; v372; v374b; weight_c	Please download patch and documentation for correcting errata as of 2011-03-15 in EVS 2008 Integrated Dataset (v. 2.0.0): ZA4800_v2-0-0_patch_1.zip; ZA4800_v2-0-0_p1_readme.doc
	2011-3-15	v106_cs v108_cs v264 v265 v305b v307b v310b v312b v343b v368b_CC v336_cs v344_cs v355_csv353W_cs v353M_cs v353Y_cs v368b_N3 v371b_N3 v368b_N2 v371b_N2 v368b_N1 v371b_N1 v376	Correction of value labels with country specific characters: Please download the Unicode patch_2 for correcting the labels in the Integrated Dataset (v. 2.0.0): ZA4800_v2-0-0_patch_2.zip
	2011-3-15	v1 to v5	Correction of the order of variables v1 to v5 in the Swedish data set: v1=v2, v2=v3, v3=v4, v4=v5, v5=v1.
	2011-3-15	v106	In the Norwegian data set hindu is coded as ´5: muslim´, but should be ´6: hindu´.
	2011-3-15	v106_cs, v108_cs	Correction of value label of country specific code 498096 and addition of missing value label of code 499001.
	2011-3-15	v136 to v147	Change of value labels of v136 to v147 into 1 "very important" 2 "rather important" 3 "not very important".
	2011-3-15	v284 to v294	Notification of deviant question wording of Q83 and Q84 . The phrase "feel concerned about" has been translated differently in several field questionnaires, for instance, in some cases it has been translated into "worried about", in other cases as "involved in".
	2011-3-15	v308	Illogical answer pattern: In 27 cases (AZ, HR, NCY, FR, DE, LV, LU, MD, SK, SI, ES, UA) is the year of birth of respondent > year in which respondent came to live in [country].
	2011-3-15	v322	Illogical answer pattern: In 110 cases (BE, HR, CZ, FI, FR, DE, IS, IE,IT, RO, SK, SI, ES, SE, TR, UA, MK, GB, NIR) is the year of birth of respondent > year in which firstborn child was born.
	2010-9-10	v46 to v60	Notification of deviant answer pattern in item battery of Q6 in Northern Cyprus.
	2011-3-15	v353m_pp	V353m_ppp was calculated with the CS income variable. In Bulgaria some respondents had score 0, so they got 0 on v353m_pp. To

Abbildung 5: Versionshistorie im DBK

Eine weitere Standardisierung der Angaben wurde durch die Einführung einer kontrollierten Liste für Untersuchungsgebiet nach ISO3166-1 und ISO3166-2 für Nationen und sub-nationale Einheiten erreicht. Ebenfalls standardisiert wurden die Erhebungszeiträume im Format TT.MM.JJJJ kompatibel zu ISO 8601 und

der Möglichkeit Zeiträume (von-bis) anzugeben. Eine Standardisierung der bisher im Freitext erfassten Literaturangaben zu den Studien wird zurzeit durchgeführt.

Der GESIS Datenbestandskatalog enthält ebenfalls Links zum Datenzugang. Seit Anfang 2012 können viele Studien der Zugangsklasse A (frei für wissenschaftliche Verwendung) direkt im DBK über einen Download erreicht werden. Alle weiteren Studien können über ein Warenkorb-System bestellt werden. Dazu ist für Nutzer lediglich eine kostenfreie Registrierung mit Angabe des Forschungsprojekts nötig.

Standardisierung

Eine standardisierte Dokumentation, wie sie mit DDI möglich ist, erlaubt den einfachen Austausch von Metadaten und Daten zwischen den Akteuren im Forschungsdaten-Lebenszyklus. Sie führt zu einer einfacheren Übernahme in neue Systeme und Anwendungen, so dass die Wiederverwendung der Dokumentation oder einzelner Teile möglich wird. Die Standardisierung führt auch zu klareren Bedeutungen einzelner Teile der Dokumentation. Daher ist die Standardisierung für die langfristige Sicherung von Forschungsdaten und ihre Nachnutzung unerlässlich. Die Etablierung eines Standards kann natürlich nur in der Community erreicht werden, die die verwendeten Dokumentationen benutzt. Eine Standardisierung erfordert daher zunächst einen höheren Aufwand bei der Dokumentation in allen Phasen des Lebenszyklus. Darüber hinaus ist eine Standardisierung ein dauernder Prozess, da sich durch Weiterentwicklungen neue Anforderungen ergeben. Der Mehrwert durch eine Standardisierung wiegt diese Anstrengungen aber mehr als auf.

Literatur

Bauske, F. (1992): Europäische Informationsbasis über Datensätze in CESSDA-Archiven. ZA-Information 31, 109-111.

Bauske, F. (2000): Das Studienbeschreibungsschema des Zentralarchivs. ZA-Information 47, 73-80.

Blank, G. and Rasmussen, K. B. (2004): The Data Documentation Initiative. The Value and Significance of a Worldwide Standard. Social Science Computer Review 22 (3), 307-318.

Consultative Committee for Space Data Systems (2002): Reference model for an Open Archival Information System (OAIS). CCSDS 650.0-B-1 Blue Book, Januar 2002.

Gregory, A./Heus, P. and Ryssevik, J. (2010): Metadata. In: German Data Forum (RatSWD) (Ed.): Building on Progress. Expanding the Research Infrastructure for the Social, Economic, and Behavioral Sciences. Opladen & Farmington Hills, 487-508.

Hausstein, B. und Zenk-Möltgen, W. (2011): da|ra – Ein Service der GESIS für die Zitation sozialwissenschaftlicher Daten. In: Schomburg, S./Leggewie, C./Lobin, H. und Puschmann, C. (Hrsg.): Digitale Wissenschaft: Stand und Entwicklung digital vernetzter Forschung in Deutschland. Beiträge der Tagung vom 20./21. September 2010. Köln, 139-147.

Jääskeläinen, T./Moschner, M. and Wackerow, J. (2009): Controlled Vocabularies for DDI 3: Enhancing Machine-Actionability. IASSIST Quarterly 33, 34-39.

Jensen, U./Katsanidou, A. und Zenk-Möltgen, W. (2011): Metadaten und Standards. In: Büttner, S./Hobohm, H. und Müller, L. (Hrsg.): Handbuch Forschungsdatenmanagement. Bad Honnef: Bock u. Herchen, 83-100.

Kramer, S./Oechtering, A. und Wackerow, J. (2009): Data Documentation Initiative (DDI): Entwicklung eines Metadatenstandards für Forschungsdaten in den Sozialwissenschaften. KIM-Technology Watch Report, September 2009.

Mochmann, E. (1979): Bericht über die IASSIST Konferenz in Ottawa. ZA-Information 4, 24-27.

Vardigan, M./Heus, P. and Thomas, W. (2008): Data Documentation Initiative: Toward a Standard for the Social Sciences. The International Journal of Digital Curation 3 (1), 107-113.

Zenk-Möltgen, W. und Habbel, N. (2012): Der GESIS Datenbestandskatalog und sein Metadatenschema. Version 1.8. Köln: GESIS Technical Reports 2012/01.

N.N. (1990): Neuauflage des Datenbestandskatalogs des Zentralarchivs. ZA-Information 27, 5-15.

Reinhard Altenhöner und Claudia Oellers (Hrsg.):

LANGZEITARCHIVIERUNG VON FORSCHUNGSDATEN

STANDARDS UND DISZIPLINSPEZIFISCHE LÖSUNGEN

Scivero Verlag 2012
ISBN 978-3-944417-00-4
16,70 €

Die Langzeitarchivierung von Forschungsdaten ist eine Voraussetzung für gute wissenschaftliche Praxis. Sie weist drei zentrale Bereiche auf: die Dokumentation der Forschungsdaten, deren langfristige Aufbewahrung sowie die Bereitstellung eines Zugangs zu den Daten. Ohne diese infrastrukturellen wie organisatorischen Voraussetzungen sind die Daten für die wissenschaftliche Sekundärnutzung, also für die Überprüfung von Ergebnissen und auch für die Beantwortung neuer Forschungsfragen nur eingeschränkt verfügbar. Das vorliegende Buch gibt einen Überblick über bestehende Standards und liefert einen Beitrag zur Diskussion über Voraussetzungen zur Archivierung von Datenbeständen. Es ist somit gleichermaßen für Infrastruktureinrichtungen, Fachbibliotheken, Archive, Wissenschaftler und alle, die im weitesten Sinne mit der Verfügbarmachung von Forschungsdaten betraut sind, lesenswert.

Bestellungen können über unsere Website www.ratswd.de/buecher aufgegeben werden.

SCIVERO